User Facilities for 21st Century Systems Biology: Providing Critical Technologies for the Research Community

Office of Science
November 2002
DOEGenomesToLife.org

Executive Summary

Realizing the Potential of the Genome Revolution

The revolution in biology triggered by the Human Genome Project promises far-reaching benefits to our nation and environment. Today, scientists have in hand the complete DNA sequences of genomes for many organisms—from microbes to plants to humans. This knowledge makes it possible to address the ultimate goal of modern biology: to achieve a fundamental, comprehensive, and systematic understanding of life. This goal is founded, as is life itself, on the genome, which contains the basic information necessary for the construction and operation of a living organism. The new Genomes to Life (GTL) program combines advanced technologies with the information found in the DNA of microbial genomes to establish a foundation for achieving this goal. Obtaining a deep level of knowledge about the diverse natural capabilities of microbes will allow scientists, both in GTL and the broader scientific community, to use those capabilities to help solve challenges in energy security, environmental cleanup, and global climate change.

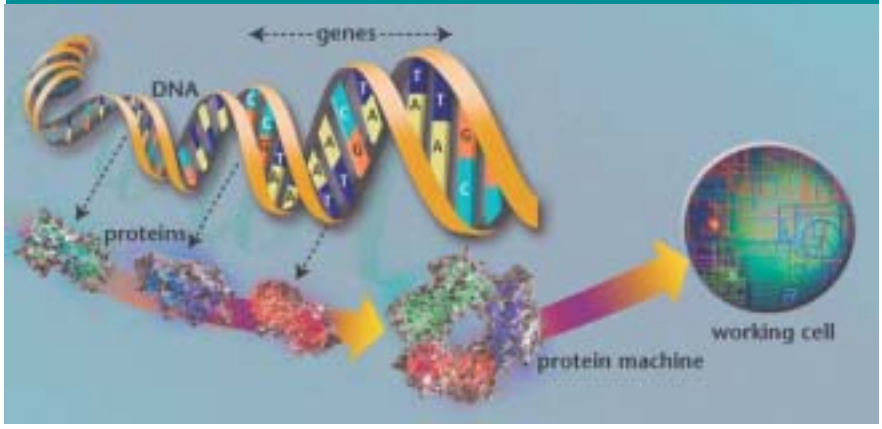
GTL scientific goals target the fundamental processes of living systems by studying them on three levels: (1) proteins and multicomponent molecular assemblies ("machines") that perform most of the cell's work, (2) gene regulatory networks that control these processes, and (3) microbial associations or communities in which groups of cells carry out the processes in nature. These tasks will require advanced experimental and computational methods and capabilities to assimilate, understand, and model the data on the scale and complexity of real living systems and, in the process, to build a dynamic knowledge base from this information. The resulting GTL knowledge base will provide data, models, and simulations of expression, pathways, and network systems, molecular machines, and cell and community processes for the entire research community.

Genomes to Life is designed to support the launching of biology onto a new trajectory—one that will empower biologists to pursue completely new approaches to discovery, based on explorations of whole functioning systems instead of the more traditional focus on cellular components. Genes encode proteins, which work together to carry out most of the activities in the cell (see figure below). These processes are orchestrated dynamically in an intricate labyrinth of pathways and networks that make the cell "come alive." Understanding cellular activities in a realistic context—known as systems biology—ultimately will transform biology to a more quantitative and predictive science that will enable effective and economical solutions to many of DOE's most pressing challenges. These capabilities will inspire revolutionary solutions to DOE mission challenges and transform the entire life sciences landscape, from agriculture to human health.

The Information Challenge

Thousands of microbes have capabilities of interest (see "Why Microbes," p. 4), and each microbial genome contains thousands of genes capable of

Genomes to Life : From DNA Sequence to Living Systems



GTL User Facilities Hallmarks

Four Enabling User Facilities

The technologies needed for systems biology require economies of scale achievable only at major facilities. To meet this challenge, the DOE Office of Biological and Environmental Research (BER) and Office of Advanced Scientific Computing Research (OASCR) propose to develop a powerful new core consisting of four complementary user facilities for both GTL and the broader scientific communities. Each will build on the capabilities of the others, moving from the use of genomic data to systematically identify, produce, and characterize microbial proteins and fully decipher how cells use the proteins to carry out life processes.

producing an even-greater number of proteins. Much biological research is now centered on completely characterizing proteins and their higher-order structures (sometimes called molecular machines) and relating that information to genome sequence. Molecular machines carry out chemical reactions, generate mechanical forces, transport metabolites and ions, and make possible every action of a biological system. Genomes contain regulatory elements that coordinate protein production and molecular machine assembly and are themselves cued by signals from the environment, including other microbial populations in their ecological community. A systems approach thus must extend from each genome throughout the population or community, encompassing thousands of proteins, molecular machines, pathways, networks, cells, and, eventually, their cellular systems and environments.

Just as DNA sequencing capability was completely inadequate at the beginning of the Human Genome Project, the quantity of data that must be collected and analyzed for systems-biology research far exceeds current capabilities and capacities. Collecting and using such data and reagents will require coordination and integration of dozens of high-throughput technologies and approaches, some not yet refined or even developed. The recent availability of genome data, emerging technologies, and high-performance computing and informatics tools and technologies now make such an approach practicable.

The facilities will make possible new avenues of inquiry, fundamentally changing the course of biological research and greatly accelerating the pace of discovery. Numerous piloting activities funded primarily by BER have established a foundation for future work and underscored the need for advanced technology user facilities accessible by the whole biological research community. Projects include systems biology and development of advanced technologies. To provide the powerful resources needed, the new facilities will use unique, high-throughput combinations of state-of-the-art instrumentation and technologies, automation, and tools. These time-phased facilities will be optimized and developed concurrently with research programs and integral computing and information infrastructure and tools. A brief description of each facility follows.

Facility I: Production and Characterization of Proteins

A microbe's genome contains instructions for making the proteins that perform nearly all the functions of life, including those that can contribute to DOE missions such as energy generation, environmental cleanup, and carbon sequestration. If we can understand how proteins "do their work," we can use them to help solve these and many other problems.

Facility I will use highly automated processes to mass-produce and characterize proteins directly from genome information and affinity reagents ("tags") to identify, capture, and monitor the proteins.

Facility II: Whole Proteome Analysis

All the proteins encoded in the genome make up an organism's "proteome." The cell does not generate all these proteins at once but rather the set required at a particular time to produce those functions dictated by environmental cues and the organism's life strategy. To make use of any microbe's capabilities, we must understand the principles of these processes.

Facility II will characterize the expressed proteomes of diverse microbes under different environmental conditions as an essential step toward determining the functions and interactions of individual proteins and sets of proteins.

Facility III: Characterization and Imaging of Molecular Machines

Cells are biological "factories" that perform and integrate thousands of discrete and highly specialized processes through the coordinated use of molecular "machines" composed of assemblies of proteins and other molecules.

Facility III will isolate, identify, and characterize thousands of molecular machines from microbes and develop the ability to image component proteins within complexes and to validate the presence of the complexes within cells.

Facility IV: Analysis and Modeling of Cellular Systems

The final step in achieving a comprehensive understanding of living systems will require the ability to measure and predict dynamic events within individual cells.

Facility IV will combine advanced computational, analytical, and experimental capabilities for the integrated observation, measurement, and analysis of the spatial and temporal variations in the state of cellular systems—from individual microbial cells to complex communities and multicellular organisms.

Office of Science—At the Forefront of the Biological Revolution

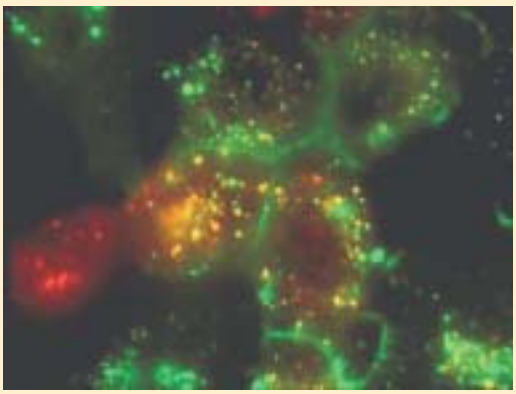
The knowledge and resources generated by scientists using the new facilities will provide GTL and the scientific community a powerful set of tools for conducting systems biology science, resulting in unique insights and opening new fields of scientific inquiry. These user facilities also will promote cross-disciplinary education of the first generation of scientists fully trained in systems biology and will attract faculty, postdocs, and students to GTL research.

Effective use of microbial and other biological systems and components will generate new biotechnological industries involving fuels, biochemical processing, nanomaterials, and broader environmental and biomedical applications.

The Office of Science has the capabilities and institutional traditions to bring the biological, physical, and computing sciences together at the scale and complexity required for success. Its academic affiliations, national laboratories, and other resources include major facilities for DNA sequencing and molecular-structure characterization, OASCR's high-performance computing resources, the expertise and infrastructure for technology development, and a tradition of productive multidisciplinary research essential for such an ambitious and complex program. In the effort to understand biological systems, these strong assets and the GTL program will complement and extend the capabilities and efforts of research supported by the National Institutes of Health, National Science Foundation, other agencies and institutions, and industry.

Lighting Up Proteins in Cells

Fluorescent labeling provides a way to study the functions of specific proteins in living cells. It allows direct observation in real time of the protein's location in time and space as well as the interactions of proteins with each other and with other cellular elements. The thousands of affinity reagents produced by automated methods in Facility I (see p. 2 and more detailed facility description beginning on p. 10) will provide scientists a means to study these components in living cells, both in GTL user facilities and in the scientists' own labs. For more information on this image, which shows the distribution and expression levels of two proteins in a grouping of cells, see p. 12.



Why Microbes?

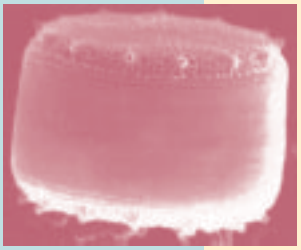
The ability of this planet to sustain life is largely dependent on microbes. They are the foundation of the biosphere, controlling earth's biogeochemical cycles and affecting the productivity of the soil, quality of water, and global climate. Microbial research is one of the most exciting frontiers in biology today, revealing the hidden architecture of life and the dynamic, life-sustaining processes on earth. Achieving an understanding of these secrets promises revolutionary solutions to many currently intractable energy and environmental problems.

Microbes must recognize available sources of energy to survive and thus have become masters at harvesting it in almost any form. Optimized to capture energy and materials for growth and cell maintenance, microbial cells can mitigate environmental threats such as toxins or extremes in pH, temperature, and salinity. Their life strategies enable microbes to carry out sophisticated biochemical functions for degrading wastes and organic matter, cycling nutrients, and, as part of the photosynthetic process, converting sunlight into energy and "fixing" (storing) CO_2 from the atmosphere.

The diversity and range of their environmental adaptations mean that microbes long ago "solved" many problems for which scientists are seeking solutions today. DOE's science missions require novel approaches for cost-efficient environmental cleanup, production of fuels (e.g., methane, ethanol, and hydrogen), and mitigation of global climate change. Microbes are capable of carrying out all these processes, but fully harnessing their natural capabilities first will require a complete understanding of their biological systems, the ultimate goal of Genomes to Life.

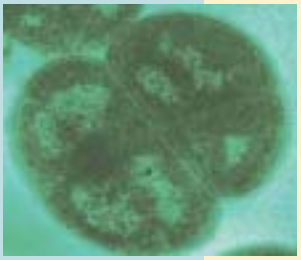
Microbes for DOE Missions

Carbon Sequestration



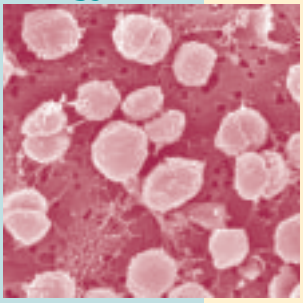
Thalassiosira pseudonana: Ocean diatom that is major participant in biological pumping of carbon to ocean depths.

Bioremediation



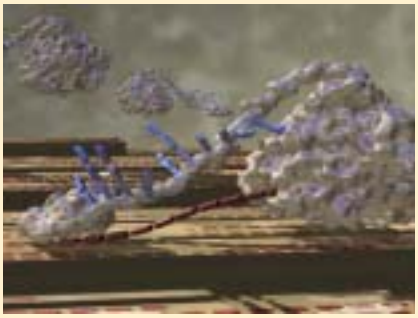
Deinococcus radiodurans: Survives extremely high levels of radiation and has high potential for radioactive waste cleanup.

Energy Production

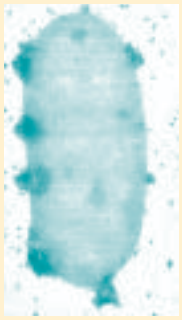


Methanococcus jannaschii: Produces methane; contains enzymes that withstand high temperatures and high pressure, possibly useful for industrial processes.

Cellulose Degradation



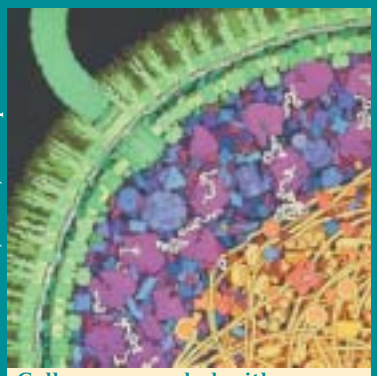
Cellulase molecular machine



Microbulbifer

Structures on the surface of *Microbulbifer* contain molecular machines (similar to cellulase, pictured above) that can break down cellulose in plant cell walls, an important step in converting cellulose to ethanol. See the description of Facility III: Characterization and Imaging of Molecular Machines to learn more about the science that can help enable this important energy application (p. 22).

Analyzing Microbes Requires Economies of Scale



D. Goodsell, 1999, with permission

Cells are crowded with many components, including proteins. A flagellum (green) protrudes from the surface. Note the molecular machine (also green) at the base that produces the flagellum's movement.

Genomes to Life will begin the analysis of a single biological system such as a microbe by using sequence databases to generate and study thousands of microbial proteins and the tags that allow them to be identified, captured, characterized, manipulated, and imaged in living systems. These resources will enable measurement of the spectrum of individual proteins, their assemblies ("molecular machines"), and associated metabolites that occur within cells and cellular communities under different experimental conditions. Analysis and modeling of cellular systems will combine knowledge of pathways, networks, and molecular machines to generate understanding of cellular and multicellular systems. High-performance computing will be used to assimilate and integrate the resulting data and information.

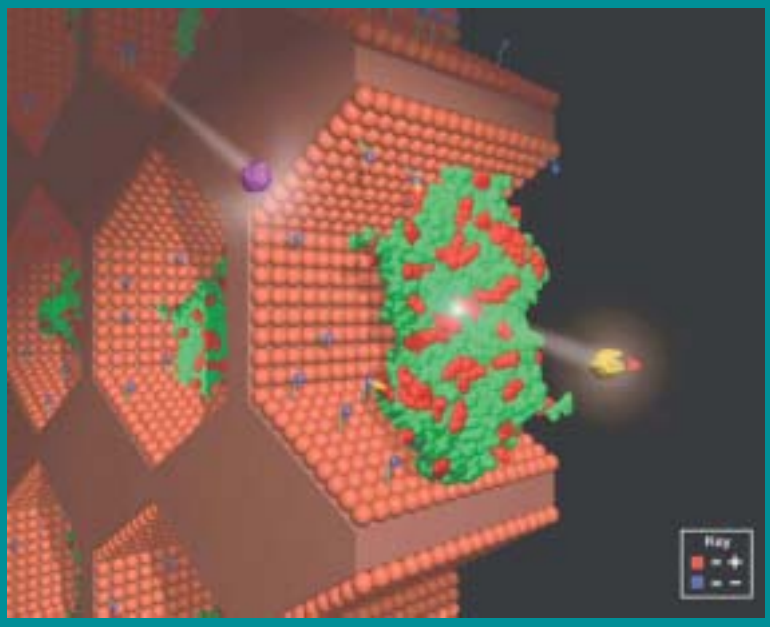
Large-Scale Facilities Spur Cost, Productivity Improvements

The dramatically increased productivity and reduced costs achieved in the Human Genome Project via high-throughput sequencing facilities provide the paradigm for the dedicated industrial-scale facilities envisioned for Genomes to Life.

Many elements of this analysis are practical only in large facilities. BER and OASCR propose the establishment of four high-throughput, integrated core user facilities to develop and implement these technologies (see text for details). These resources will create an integrated knowledge base for systems-biology science available to researchers worldwide. Applications of this level of knowledge will be far-reaching across the life sciences.

A Possible Application of Knowledge Gained in GTL Facilities

Harnessing Enzymes to Inactivate Contaminants, Generate Energy, Sequester Carbon



Miniaturization technologies can be combined with biological components such as enzymes to create novel systems that use an array of microbial processes but do not require living cells. In this figure, the enzyme organophosphorus hydrolase (OPH) is embedded in a synthetic nanomembrane (mesoporous silica) that enhances its activity and stability [*J. Am. Chem. Soc.* **124**, 11242–43 (2002)]. Applications such as this could enable development of efficient enzyme-based ways to produce energy, remove or inactivate contaminants, and sequester carbon to mitigate global climate change. It also could be highly useful in food processing, pharmaceuticals, separations, and the production of industrial chemicals. For more details on this illustration and large-scale protein production in GTL, see Facility I: Production and Characterization of Proteins, p. 10.

Genomes to Life Program Management

Pilots and Awards

Over the past several years, BER has funded pilot studies in technology development and systems biology. These projects have demonstrated mass spectrometric analysis of microbial proteomes, development of new imaging modalities, small-scale generation of microbial proteins, and development of computational tools for first-generation genome analysis and annotation. These pilots are producing data and experience to identify bottlenecks, areas for technology development, and scale of facilities needed.

In July 2002, DOE announced five major research awards for GTL systems biology to consortia involving many institutions (see below) and totaling $103 million over the next 5 years. These awards represent the culmination of nearly 3 years of planning by the DOE Office of Science and hundreds of scientists at universities, national laboratories, and industry. The microbes studied in the pilot projects as well as the 2002 awards have potential for bioremediating metals and radionuclides, degrading organic pollutants, producing hydrogen, sequestering carbon, and demonstrating importance in ocean carbon cycling. All have had their genetic sequences determined under DOE's Microbial Genome Program.

2002 Awards

- Oak Ridge National Laboratory and Pacific Northwest National Laboratory. "Genomes to Life Center for Molecular and Cellular Systems: A Research Program for Identification and Characterization of Protein Complexes"
- Lawrence Berkeley National Laboratory. "Rapid Deduction of Stress Response Pathways in Metal/Radionuclide–Reducing Bacteria"
- Sandia National Laboratories. "Carbon Sequestration in *Synechococcus*: From Molecular Machines to Hierarchical Modeling"
- University of Massachusetts, Amherst. "Analysis of the Genetic Potential and Gene Expression of Microbial Communities Involved in the In Situ Bioremediation of Uranium and Harvesting Electrical Energy from Organic Matter"
- Harvard Medical School. "Microbial Ecology, Proteogenomics, and Computational Optima"

Other Participating Institutions

Argonne National Laboratory

Brigham and Women's Hospital

Diversa Corporation

Los Alamos National Laboratory

Massachusetts General Hospital

Massachusetts Institute of Technology

National Center for Genome Resources

The Institute for Genomic Research

The Molecular Science Institute

University of California (Berkeley, San Diego, Santa Barbara)

University of Illinois

University of Michigan

University of Missouri

University of North Carolina

University of Tennessee (Knoxville, Memphis)

University of Utah

University of Washington

URL for Call for Proposals: www.er.doe.gov/production/grants/Fr03-05.html

GTL Web Site: DOEGenomesToLife.org

Fostering Research Community Participation

The Web site of the Genomes to Life program is designed to inform and foster participation in this exciting new undertaking by multidisciplinary investigators in the greater scientific community, science administrators, related policymakers, educators, and the general public. A suite of educational resources such as genome posters and handouts is accessible, and teachers can request multiple copies of materials. All GTL publications are posted, as are downloadable images, workshop reports, funding announcements, and abstracts of cutting-edge technologies.

Direct Web Access

- doegenomestolife.org/pubs.html
- doegenomestolife.org/gallery/images.html
- doegenomestolife.org/research/index.html
- www.ornl.gov/hgmis/education/education.html

Achieving a Molecular-Level Understanding of Life: A National Science Priority

OMB, DOE Budget Statements

The recent Office of Management and Budget's (OMB) Office of Science and Technology Policy memo, "FY 2004 Interagency Research and Development Priorities," declares that achieving a "molecular-level understanding of life processes" is a national science priority. The memo notes, "Sequence and structure data, coupled to modern computational power and to our ability to manipulate biological systems at the molecular level, will yield new experimental approaches that have the potential to unravel the complexity of life at the molecular, cellular, and organismal levels."

Solving DOE mission challenges requires the application of a systems approach to biology. DOE has the capabilities and institutional traditions to bring the biological, physical, and computing sciences together at the scale and complexity required for success in these efforts. The DOE FY 2003 budget request noted that "one of the most exciting areas of exploration is in the study of microbes—'bugs' that withstand extreme environments and may one day solve our energy-production problems and eat their way through our toughest environmental-cleanup areas."

AAM Recommends New Technologies

Specific recommendations from the American Academy of Microbiology's (AAM) 2001 colloquium report on "Microbial Ecology and Genomics: A Crossroads of Opportunity," are (1) "Develop new technologies, including methods for measuring the activity of microorganisms (at the level of populations and single cells); approaches to cultivating currently uncultivable species; and methods for rapid determination of key physiological traits and activities; and (2) "Establish mechanisms to encourage the necessary instrument development." A related recommendation is to (3) "Encourage instrumentation development through collaboration with device engineers, chemists, physicists, and computational scientists, since uncovering the diversity and activities of the microbial world is dependent on such advances."

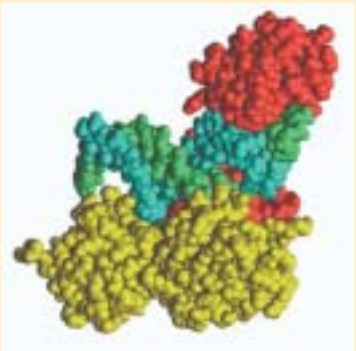
Another recommended goal was to (4) "Develop technology and analysis capability to study microbial communities and symbioses holistically, measuring system-wide expression patterns (mRNA and protein) and activity measurements at the level of populations and single cells."

BERAC Approves GTL Facilities Strategy

Since the inception of the GTL program, the DOE Offices of Advanced Scientific Computing Research (OASCR) and Biological and Environmental Research (BER) have sponsored 15 workshops over the last 2 years to guide program implementation. Participants represented the breadth of the scientific community in industry, national laboratories, and academia. A strategic plan for developing facilities to serve the entire community was approved by the BER Advisory Committee (BERAC) in April 2002. The GTL roadmap is being updated in 2002.

Understanding Molecular Machines Requires Ultrascale Computing

Molecular machines—assemblies of proteins and other chemical components—carry out most of life processes. Computationally simulating molecular machine activity is a critical prelude to understanding and using microbial capabilities and requires levels of computing power beyond that available today. For more details on computational modeling of molecular machines, see sidebars, pp. 26 and 33.



User Facility Governance

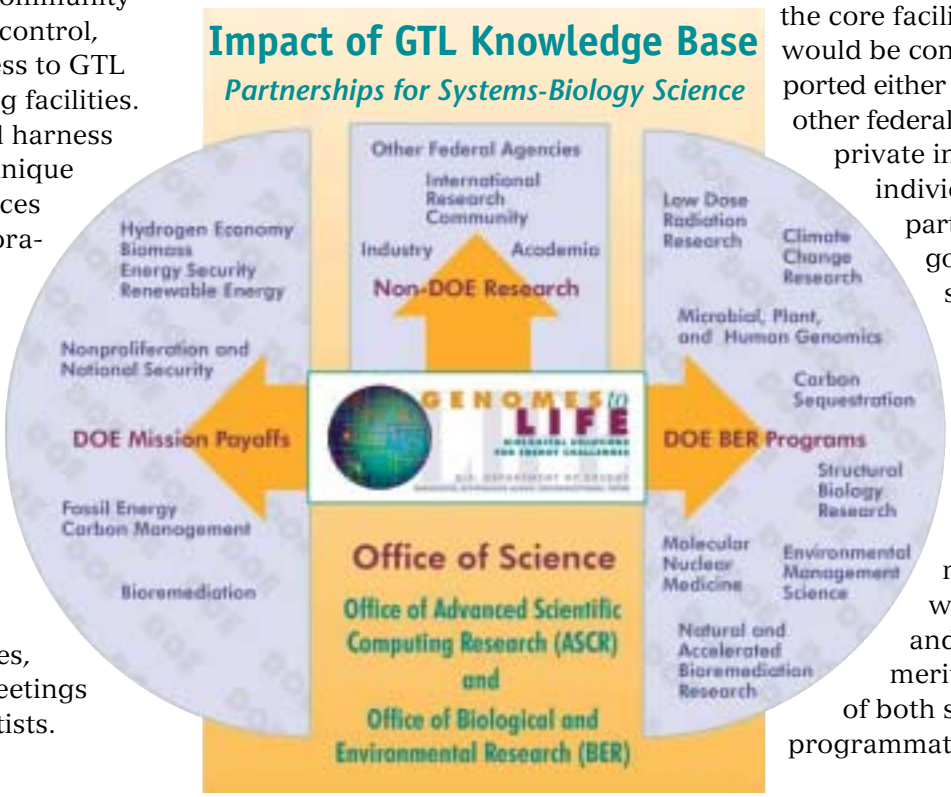
The goal of the GTL program is to explore the amazingly diverse characteristics and capabilities of microbes, gain insight into the life processes of cells and communities of cells, and understand functions that can be exploited for solutions to mission problems. A consensus is growing within the scientific community that achieving a higher level of biological understanding requires a new paradigm based on an integrated or systems approach—a synergistic application of experiment, theory, and modeling.

This new paradigm will require numerous and significantly more complex approaches and dedicated user facilities that provide the broader scientific community with technologies and computing and information infrastructure to gain the necessary innovation, efficiency, and efficacy. Biology will be democratized, and the most sophisticated and comprehensive capabilities, reagents, and data will be available to investigators lacking such integrated technology suites in their own laboratories or institutions. Moreover, new avenues of principal investigator–driven research will be enabled. Some of the GTL program's capabilities and facilities will transcend the program and will directly and indirectly benefit other public and private programs in systems biology.

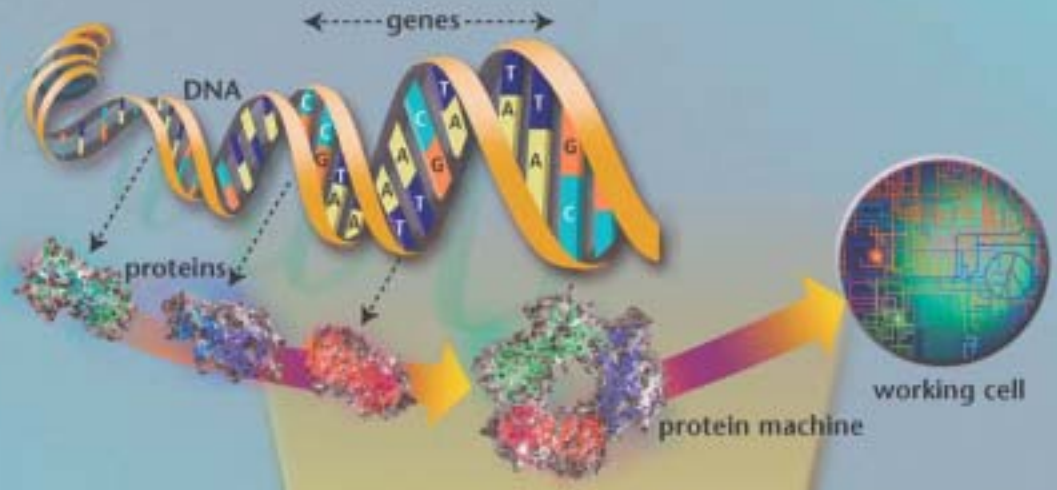
A sound governance and access model will ensure that the scientific community maintains positive control, influence, and access to GTL resources, including facilities. This enterprise will harness and integrate the unique powers and resources of the national laboratories, academia, and industry. DOE will seek advice from the scientific community through the usual mechanisms that could include the National Academy of Sciences, inter- and intra-agency advisory committees, workshops, and meetings of supported scientists.

Management and financing of programs have evolved over the years, particularly for facilities, and most user facilities are now managed with what is termed the "steward-partner model." This model was developed to ensure that user facilities provide the maximum scientific benefit to the broadest possible research community in the most cost-effective manner. It was implemented in the report, *Synchrotron Radiation for Macromolecular Crystallography*, Office of Science and Technology Policy (February 5, 1999). The model is described in some detail in the National Research Council report, *Cooperative Stewardship: Managing the Nation's Multidisciplinary User Facilities for Research with Synchrotron Radiation, Neutrons, and High Magnetic Fields* (National Academy Press, 1999). It also is followed in the recent report, *Office of Science and Technology Policy Interagency Working Group on Neutron Science: Report on the Status and Needs of Major Neutron Scattering Facilities and Instruments in the United States* (June 2002). (Reports: www.ostp.gov/Science/html/cassman_rpt.html, www.nap.edu/books/0309068312/html/index, and www.ostp.gov/html/NeutronIWGReport.pdf, respectively.)

Investment in GTL facilities would be a national commitment to enable relevant cutting-edge science that also supports DOE programs. In this model, DOE (the steward) would manage and fund the core facilities. Research would be conducted and supported either by the steward or by other federal agencies, industry, private institutions, or individual scientists (the partners). Principles governing the steward-partner model would be used to provide GTL facility resources to the scientific community. Good stewards of the investments and trust would determine use and access by objective merit-based peer review of both scientific quality and programmatic relevance.



From DNA Sequence to Living Systems: Advanced Technology Facilities Critical to Genomes to Life



High-Throughput Technologies

Facility I: Production and Characterization of Proteins

- Use genome data to generate and characterize proteins, tags, and other resources needed to study microbes
- In vivo, in vitro, and chemical synthesis of proteins
- Multiple biophysical characterizations
- Computational tools for tracking and biophysical analysis
- Automated superannotation of genome data

Facility II: Whole-Proteome Analysis

- Measure proteome and metabolites for a cell or community systems under controlled conditions
- Gain functional insights by characterizing known and unknown dynamic processes to correlate proteins and machines that work together in a process
- Large-scale proteome analysis via mass-spectrometry
- High-capacity cultivation systems
- Optical analysis and cell sorting
- Computational tools for petabyte-scale data

Facility III: Characterization and Imaging of Molecular Machines

- Isolate the repertoire of molecular machines
- Characterize machines in terms of composition and molecular organization
- Biophysical and imaging characterization of molecular machines
- Biophysical and imaging tools
- Computational tools for molecular machine modeling and image analysis

Facility IV: Analysis and Modeling of Cellular Systems

- Couple knowledge of pathways, networks, and molecular machines to generate understanding of cellular and multicellular systems
- Measure structure and properties of a single cell in a population or community under controlled conditions
- Molecular machine imaging and characterization in living cells
- Culture and analysis of complex microbial communities
- Computational tools and algorithms for modeling cells

- **Comprehensive understanding of living systems**
- **Applications to DOE missions and across the life sciences**

Facility I: Production and Characterization of Proteins

A microbe's genome contains instructions for making the proteins that perform nearly all the functions of life, including those that can contribute to DOE missions such as energy generation, environmental cleanup, and carbon sequestration. If we can understand how these proteins "do their work," we can use them to help solve these and many other problems.

GTL Facility I will use highly automated processes to mass-produce and characterize proteins directly from genome information and affinity reagents ("tags") to identify, track, quantify, manipulate, capture, and monitor the proteins.

Strategic Intent

Virtually every cellular chemical reaction and physical function necessary for sustaining life is controlled and mediated by proteins generally organized into multiprotein complexes, or "molecular machines." High-throughput, automated protein and affinity-tag production and subsequent functional analysis of proteins and complexes will enable the study of chemical and physical interactions of proteins that underlie biology. A typical microbial genome has 2000 to 5000 genes that contain both the recipes for the production of thousands of proteins and the regulatory signals that control production. The cell does not generate all the proteins at once but rather the set required at a particular time for the functionality dictated by environmental cues and the organism's life strategy.

A systems-level understanding of cellular behavior will require experimental data for a significant portion of an organism's proteins. We must have the capability to produce essentially all the thousands of proteins encoded in each of many different genomes.

While selection of particular proteins may be simplified by comparative genomic analyses, production rates for proteins and their affinity tags will be many times higher than currently available. The associated characterization data for these reagents will better benefit the scientific community when it is generated under automated, high-throughput, and standardized conditions because it will be highly reliable and reproducible. Protein availability will enable production of affinity reagents to identify the components of molecular machines and to specifically capture, label, and track proteins in living systems. Data and reagents produced in Facility I will be invaluable resources for understanding molecular machines and cellular processes.

The overall objective of the facility is to produce milligram quantities of tens of thousands of full-length, functional proteins per year; generate multiple affinity tags for each protein; provide initial biophysical characterizations of each protein; and construct a comprehensive production and characterization database. Proteins, tags, protocols, and annotations produced in this facility will be motivated by the needs of the DOE GTL research program, DOE science missions, and the broader biological research community. Expression vectors, proteins, affinity reagents, and data produced will be foundational resources for all other GTL facilities and the entire scientific community.

Specialized, large-scale facilities are needed to achieve the necessary economy of scale and output of standardized characterization data associated with each protein and affinity reagent. This capability is an essential component of the nation's science infrastructure.

Project Purpose and Justification

Protein production is currently limited by economic and technological constraints. In the absence of significant automation, costs are prohibitive for genome-scale efforts ($10,000 to $25,000 per protein plus similar costs for tags). Current efforts costing hundreds of millions of dollars by individual-investigator approaches are focused on only a fraction of selected proteins. This protein subset and associated affinity tags typically are made independently in multiple laboratories at great expense. Unfortunately, characterization data about processes that worked and those that failed usually are not standardized, so user communities have difficulty obtaining reliable and comparable data to build models. Furthermore, many data associated with production, storage, and characterization are not preserved. The goal of this facility is to address these issues and, in particular, to develop comprehensive databases for processes that will aid automation and dramatically lower the costs of producing proteins and protein-affinity tags. Development of high-throughput production and characterization using robotics integrated with